Coding and decoding in the Evolution of communication: Information richness and Referentiality

Bernat Corominas Murtra,[1] Jordi Fortuny Andreu,[2] and Ricard Solé[1,3,4]
[1]*ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain*
[2]*Centre de Lingüística Teòrica (CLT), Facultat de Lletres, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*
[3]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA*
[4]*Institut de Biologia Evolutiva. CSIC-UPF. Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain.*

One of the most basic properties of the communicative sign is its dual nature. That is, a sign is a twofold entity composed of a formal component, which we call *signal*, and a referential component, namely a *reference*. Based on this conception, we say that a referent is coded in a particular sign, or that a sign is decoded in a particular referent. In selective scenarios it is crucial for the success of any adaptive innovation or communicative exchange that, if a particular referent a is coded in a particular signal s during the coding process, then the referent a is decoded from the sign s during the decoding process. In other words the *referentiality of a signal* must be preserved after being decoded, due to a selective pressure. Despite the information-theoretic flavour of this requirement, an inquiry into classical concepts of information theory such as entropy or mutual information will lead us to the conclusion that information theory as usually stated does not account for this very important requirement that natural communication systems must satisfy. Motivated by the relevance of the preservation of referentiality in evolution, we will fill this gap from a theoretical viewpoint, by deriving the consistent information conveyed from an arbitrary coding agent A^v to an arbitrary decoding agent A^u and discussing several of its interesting properties.

Keywords: entropy, information, referentiality, consistent information

I. INTRODUCTION

Biological Systems store and process information at many different scales (Yockey, 1992). Organisms or cells react to changes in the external environment by gathering information and making the right decisions -once such information is properly interpreted. In a way, we can identify the external changes as input signals to be coded and decoded by the cellullar machinery or information processing of neural networks, and include the exchange of signals between individuals or abstract agents sharing a given communication system (Hurford, 1989; Komarova and Niyogi, 2004; Niyogi, 2006; Nowak and Krakauer, 1999).

The ability to store information to interpret the surroundings beyond pure noise is thus an important property of biological systems. An organism or abstract agent can make use of this feature to react to the environment in a selectively advantageous way. This is possible provided that, in biological systems, a communicative signal must be necessarily linked to a referential value, that is, it must have a meaningful content. As pointed out by John Hopfield:

> *Meaningful content, as distinct from noise entropy, can be distinguished by the fact that a change in a meaningful bit will have an effect on the macroscopic behavior of a system* (Hopfield, 1994).

The meaningful content of information can be understood as something additional to classical information which is preserved through generations (or by the members of a given population in a given communicative exchange) resulting in a *consistent* response to the environment (Haken, 1978).

The explicit incorporation of the referential value in the information content is, in some sense, external to classical information theory, since, roughly speaking, the standard measure of mutual information only accounts for the relevance of correlations among sets of random variables. Indeed, one can establish configurations among coder and decoder by which mutual information is maximal but the referentiality value of the signal is lost during the communicative exchange. Let us consider the following example: Suppose a system where the event *fire* is coded as the signal a, and that such a signal a is always decoded as the event *water*. Suppose, also, that the event *water* is coded as the signal b and it is always decoded as *fire*. In this system, both the coder and the decoder depict a one-to-one mapping between input and output, and the mutual information between the set of events shared by coder and decoder would be maximum. However, if we take the system as a whole, the non-preservation of any referential value renders the communication code useless.

Not surprisingly, evolutionary experiments involving artificial agents (such as robots) include, as part of the selective pressures, the consistency of signals and referents. If survival or higher scores depend on a fitness measure which requires a proper sharing of information, the final outcome of the dynamics is a set of agents using common signals to refer to the same object (Nolfi and Mirolli, 2010; Steels, 2001; Steels and Baillie, 2003). Formally, we say that the communicative sign has a dual

nature[1]: a sign would involve a pair

$$\langle m_i, s_k \rangle, \tag{1}$$

composed of a *signal*, s_i, and a *referent*, m_k. Such pair must be conserved in a consistent communicative interchange.

The problem of consistency of the communicative process was early addressed in (Hurford, 1989), through a formalism consisting in signal/referent matrices. Further works showed the suitability of such formalism, and enabled the study of the emergence of consensus driven by selective forces (Nowak and Krakauer, 1999). These studies showed that an evolutionary process could result in a shared code by a population of interacting agents. Under this framework, the existence of optimal solutions has been studied (Komarova and Niyogi, 2004), as well as the problem of the information catastrophe or *linguistic error limit* (Nowak, 2000), using evolutionary game theory involving a payoff function accounting for the average number of well-referentiated signals.

It is the purpose of this theoretical work to rigorously identify the amount of information which conserves the dual structure of a sign, i.e., the amount of *consistent information*, and to explore some of its consequences. Specifically, we evaluate the relevance of the consistent input/output pairs, assuming that the input set and the output set are equal. The study of the behaviour of the consistent information displays interesting differences with classical Shannon's mutual information.

We should properly differentiate the problem of consistency from the problem of *absolute information content* of a given signal -or, in general, mathematical object. The latter arises from the fact that, in Shannon's information theory, the information content of a given signal is computed from the relative abundance of such a signal against the occurrences of the whole set of signals. The information content of an isolated signal is not defined (or equal to zero). This is solved by the definition of the Kolmogorov Complexity (Cover and Thomas, 1991; Kolmogorov, 1965; Ming and Vitányi, 1997), which can be understood as the absolute information content of a given signal -or mathematical object. Our purpose can be embedded in Shannon's framework. Accepting the relative nature of the information content, we attack the problem of the consistency of input/output pairs.

The paper is written in a self-contained way. Thus, beyond basics of probability theory we properly introduce the concepts and the required mathematical apparatus. At the end of the paper, a case study (the classical binary symmetric channel) is described in detail.

[1] This central property of the communicative sign resembles the *duality of the linguistic sign* pointed out by first time by the Swiss linguist Ferdinand de Saussure (Saussure, 1916). According to Saussure, a linguistic sign is a psychical unit with two faces: a signifier and a signified. The former term is close to our term 'signal' and the latter to our term 'reference'. There are, though, important differences between the information-theoretical approach we are about to develop and Saussure's conception of the linguistic sign.

II. THE MINIMAL SYSTEM AND ITS ASSOCIATED INFORMATION MEASURES

In this section we define the minimal system composed of two agents able to both code and decode a set of external events.

A. The communicative system

Consider a set of (at least, two) interacting agents *living* in a shared world (Komarova and Niyogi, 2004). Agents communicatively interact through noisy channels. The description of this system is based on the probability transition matrices defining the coding and decoding processes, the probability transition matrix for the channel and the random variables associated to the inputs and outputs, which account for the successive information processing through the system formed by two agents and the noisy channel -see fig.1. The qualitative difference with respect to the classical communication scheme is that we take into account the particular value of the input and the output therby capturing the referential value of the communicative exchange. An *agent*, A^v, is defined as a pair of computing devices,

$$A^v \equiv \{\mathbf{P}^v, \mathbf{Q}^v\}, \tag{2}$$

where $\mathbf{P}^v$ is the coder module and $\mathbf{Q}^v$ is the decoder module. The shared world is defined by a random variable X_Ω which takes values on the set of events, Ω:

$$\Omega = \{m_1, ..., m_n\}, \tag{3}$$

being the (always non-zero) probability associated to any event $m_k \in \Omega$ defined by $\mu(m_k)$. The coder module, $\mathbf{P}^v$, is described by a mapping from Ω to the set $\mathcal{S}$:

$$\mathcal{S} = \{s_1, ..., s_n\}, \tag{4}$$

to be identified as the set of signals. For simplicity, here we assume $|\Omega| = |\mathcal{S}| = n$. This mapping is realized according to the following matrix of transition probabilities:

$$\mathbf{P}^v_{ij} = \mathbb{P}_v(s_j|m_i), \tag{5}$$

which satisfies the following condition:

$$(\forall m_i \in \Omega) \quad \sum_{j \leq n} \mathbf{P}^v_{ij} = 1. \tag{6}$$

The output of the coding process is described by the random variable X_s, taking values on $\mathcal{S}$ according to the probability distribution ν:

$$\nu(s_i) = \sum_{k \leq n} \mu(m_k)\mathbf{P}^v_{ki}. \tag{7}$$

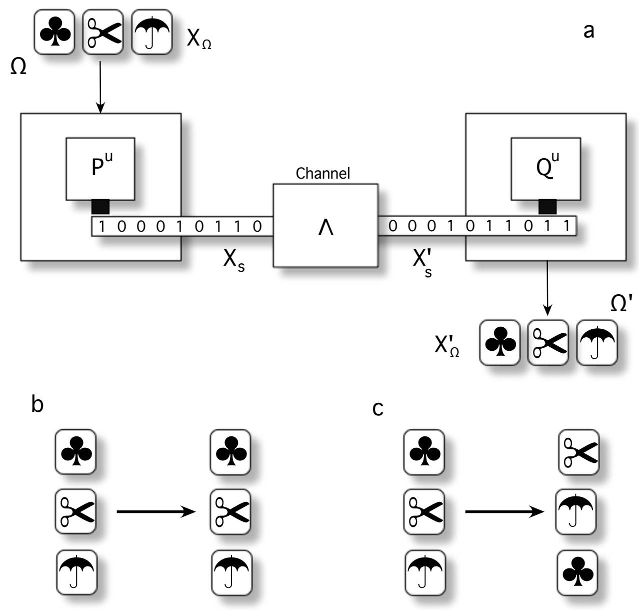

FIG. 1 The minimal communicative system to study the conservation of referentiality (a): A shared world, whose events are the members of the set Ω and whose behavior is governed by the random variable X_Ω. A coding engine, $\mathbf{P}^v$, which performs a mapping between Ω and the set of signals $\mathcal{S}$, being X_s the random variable describing the behavior of the set of signals obtained after coding. The channel, Λ, may be noisy and, thus, the input of the decoding device, $\mathbf{Q}^v$, depicted by X'_s, might be different from X_s. $\mathbf{Q}^v$ performs a mapping among $\mathcal{S}$ and Ω whose output is described by X'_Ω. Whereas the mutual information provides us a measure of the relevance of the correlations among X_Ω and X'_Ω, the *consistent information* evaluates the relevance of the information provided by consistent pairs on the overall amount of information. In this context, from a pure information-theoretical point of view, situations like b) and c) could be indistinguishable. By defining the so-called consistent information we can properly differentiate b) and c) by evaluating the degree of consistency of input/output pairs -see text.

The channel, Λ, is characterized by the $n \times n$ matrix of conditional probabilities $\mathbb{P}_\Lambda(\mathcal{S}|\mathcal{S})$, i.e.,

$$\Lambda_{ij} = \mathbb{P}_\Lambda(s_j|s_i). \tag{8}$$

The output of the composite system coder+channel, $\mathbf{P}^v\Lambda$, is described by the random variable X'_s, which takes values on the set $\mathcal{S}$ following the probability distribution ν', defined as:

$$\nu'(s_i) = \sum_k \mu(m_k)\mathbb{P}_{v\Lambda}(s_i|m_k), \tag{9}$$

where

$$\mathbb{P}_{v\Lambda}(s_i|m_k) = \sum_{j\leq n} \mathbf{P}^v_{kj}\Lambda_{ji}. \tag{10}$$

Finally, the decoder module is a computational device described by a mapping from $\mathcal{S}$ to Ω, i.e it receives $\mathcal{S}$ as the input set, emitted by another agent through the channel, and yields as output elements of the set Ω. $\mathbf{Q}^v$ is completely defined by its transition probabilities, i.e.:

$$\mathbf{Q}^v_{ik} = \mathbb{P}_v(m_k|s_i), \tag{11}$$

which satisfies the following condition:

$$(\forall s_i \in \mathcal{S}) \;\; \sum_{k\leq n} \mathbf{Q}^v_{ik} = 1. \tag{12}$$

Aditionally, we can impose another condition:

$$(\forall m_j \in \Omega) \;\; \sum_{i\leq n} \mathbb{P}_v(s_i|m_j) = 1, \tag{13}$$

which is necessary for A^v to reconstruct Ω, i.e., if the population of interacting agents share the world. By imposing condition (13) we avoid configurations in which some $m_k \in \Omega$ cannot be referentiated by the decoder agent. We notice that it is consistent with the fact that no element from Ω has zero probability to occur. Furthermore, we emphasize the assumption that, in a given agent A^v, following (Nowak and Krakauer, 1999; Plotkin and Nowak, 2000) but not (Hurford, 1989; Komarova and Niyogi, 2004) there is a priori no correlation between $\mathbf{P}^v$ and $\mathbf{Q}^v$. Finally, under the presence of another agent A^u, we can define the output of $\mathbf{Q}^v$ as the random variable X'_Ω, taking values on the set Ω and following the probability distribution μ', which takes the form:

$$\mu'(m_i) \equiv \sum_{l\leq n} \mu(m_l)\mathbb{P}_{A^u\to A^v}(m_i|m_l), \tag{14}$$

where

$$\mathbb{P}_{A^u\to A^v}(m_i|m_l) = \sum_{j,r\leq n} \mathbf{P}^u_{lj}\Lambda_{jr}\mathbf{Q}^v_{ri}, \tag{15}$$

$$\mathbb{P}_{A^v\to A^u}(m_i, m_j) = \sum_{l,r} \mu(m_j)\mathbf{P}^v_{jl}\Lambda_{lr}\mathbf{Q}^u_{ri}. \tag{16}$$

Consistently,

$$\sum_{i,l\leq n} \mathbb{P}_{A^u\to A^v}(m_i|m_l) = n. \tag{17}$$

Once we have the description of the different pieces of the problem, we proceed to study the couplings among them in order to obtain a suitable measure of the consistency of the communicative process. The first natural quantitative observable to account for the degree of consistency is the fraction of events $m_i \in \Omega$ which are consistently decoded. From eq. (16) it is straigtforward to conclude that such a fraction ($F(A^v \to A^u)$) is given by:

$$F(A^v \to A^u) = \sum_{i\leq n} \mathbb{P}_{A^v\to A^u}(m_i, m_i). \tag{18}$$

And if we take into account that the communicative exchange takes place in both directions, we have:

$$F(A^v, A^u) = \frac{1}{2}\left(F(A^v \to A^u) + F(A^u \to A^v)\right). \quad (19)$$

Putting aside slight variations, eq. (19) has been widely used as a payoff function to study the emergence of consistent codes -in terms of duality preservation- through an evolutionary process involving several agents in every generation (Hurford, 1989; Komarova and Niyogi, 2004; Nowak and Krakauer, 1999; Plotkin and Nowak, 2000). Such an evolutionary dynamics yielded important results which help understanding how selective pressures push a population of communicating agents to reach a consensus in their internal codes.

B. Mutual Information

Now we proceed to compute the mutual information among relevant variables of the system. We stress that it does not account for the referentiality of the sent signals. Instead, it quantifies, in bits, the relevance of the correlations among two random variables, as a potential message conveyer system, never specifying the referential value of any sequence or signal.

Let us briefly review some fundamental definitions and concepts of information theory. We know that, given two random variables X, Y, with associated probability functions $p(x), p(y)$, conditional probabilities $\mathbb{P}(x|y), \mathbb{P}(y|x)$ and joint probabilities $\mathbb{P}(x, y)$, its mutual information $I(X : Y)$ is defined as (Ash, 1990; Cover and Thomas, 1991; Shannon, 1948):

$$I(X : Y) \;=\; \sum_{x,y} \mathbb{P}(x, y) \log \frac{\mathbb{P}(x, y)}{p(x)p(y)}, \quad (20)$$

or equivalently:

$$I(X : Y) = H(X) - H(X|Y), \quad (21)$$

being $H(X)$ the *Shannon entropy* or *uncertainty* associated to the random variable X:

$$H(X) = -\sum_{x} p(x) \log p(x), \quad (22)$$

and $H(X|Y)$ the *conditional entropy* or *conditional uncertainty* associated to the random variable X with respect to the random variable Y:

$$H(X|Y) = -\sum_{y} p(y) \sum_{x} \mathbb{P}(x|y) \log \mathbb{P}(x|y). \quad (23)$$

We can also define the *joint entropy* among two random variables X, Y, written as $H(X, Y)$:

$$H(X, Y) = -\sum_{x,y} \mathbb{P}(x, y) \log \mathbb{P}(x, y). \quad (24)$$

A key concept of information theory is the so-called *channel capacity*, $C(\Lambda)$, which, roughly speaking, is the maximum amount of bits that can be reliably processed by the system, namely:

$$C(\Lambda) = \max_{p(x)} I(X : Y). \quad (25)$$

As usual, in our minimal system of two interacting agents we explicitely introduced the channel, Λ, as a matrix of transition probabilities between the two agents. Channel capacity is an intrinsic feature of the channel; as the fundamental theorem of information theory (Ash, 1990; Cover and Thomas, 1991; Shannon, 1948) states, it is possible to send any message of R bits through the channel with an arbitrary small probability of error if:

$$R < C(\Lambda); \quad (26)$$

otherwise, the probability of errors in transmission is no longer negligible. One should not confuse the statements concerning the capacity of the channel with the fact that given a random variable with associated probability distribution $p(x)$, we have:

$$\max I(X : Y) = H(X) = H(Y) \quad (27)$$

(provided that $C(\Lambda) > H(X)$). In those cases, we refer to the channel as *noiseless*.

Let us now return to our system. Using eq. (20) and the joint probabilities derived in eq. (16), we can compute the mutual information among X_Ω and X'_Ω when A^v is the coder and A^u the decoder, to be noted $I(A^v \to A^u)$, as follows:

$$\begin{aligned} I(A^v \to A^u) \;&=\; \sum_{j,i\leq n} \mathbb{P}_{A^v\to A^u}(m_i, m_i) \times \\ &\times \log\left(\frac{\mathbb{P}_{A^v\to A^u}(m_i, m_j)}{\mu(m_i)\mu'(m_j)}\right). \end{aligned} \quad (28)$$

Notice that, since the coding and decoding modules of a given agent are depicted by different, a priori non-related matrices, in general

$$I(A^v \to A^u) \neq I(A^u \to A^v). \quad (29)$$

The average of shared information among agent A^v and A^u will be:

$$\langle I(A^v, A^u)\rangle = \frac{1}{2}(I(A^v \to A^u) + I(A^u \to A^v)). \quad (30)$$

Clearly, since the channel is the same in both directions of the communicative exchange, the following inequality holds:

$$\langle I(A^v, A^u)\rangle < C(\Lambda). \quad (31)$$

In the next section we investigate the role of the *well*-correlated pairs and its impact in the overall quantity of information.

III. CONSISTENT INFORMATION

To obtain the amount of consistent information shared among A^u and A^v, we must find a special type of correlations among X_Ω and X'_Ω. Specifically, we are concerned with the observations of both coder and decoder such that the input and the output are the same element, i.e., the fraction of information that can be extracted from the observation of all consistent pairs $\mathbb{P}_{A^v\to A^u}(m_i, m_i)$. This fraction is captured by the so-called *referential parameter*, and its derivation is the objective of the next subsection.

A. The Referential parameter

The mutual information among two random variables is obtained by exploring the behavior of input/output pairs, averaging the logarithm of the relation among the actual probability to find a given pair and the one expected by chance. Consistently, the referential parameter is thus obtained by averaging the fraction of information that can be extracted by observing consistent pairs against the whole information we can obtain by looking at all possible ones.

1. Derivation of the Referential Parameter σ

Following the standard definitions of the information conveyed by a signal (Shannon, 1948), the information we extract from the observation of a pair input-output m_i, m_j is:

$$-\log \mathbb{P}_{A^v\to A^u}(m_i, m_j). \qquad (32)$$

Following eq. (24), the average of information obtained from the observation of pairs will be precisely the joint entropy between X_Ω and X'_Ω, $H(X_\Omega, X'_\Omega)$:

$$-\sum_{i,j\leq n} \mathbb{P}_{A^v\to A^u}(m_i, m_j)\log \mathbb{P}_{A^v\to A^u}(m_i, m_j).$$

Let us simplify the notation by defining a matrix $\mathbf{J}$. The elements of such a matrix are the joint probabilities, namely:

$$J_{ij} \equiv \mathbb{P}_{A^v\to A^u}(m_i, m_j). \qquad (33)$$

From the above matrix, we can identify the contributions of the consistent pairs by looking at the elements of the diagonal. The relative impact of consistent pairs on the overall measure of information will define the *referential parameter* associated to the communicative exchange $A^v \to A^u$, to be indicated as $\sigma_{A^v\to A^u}$. This is our key definition, and its explicit form will be:

$$\sigma_{A^v\to A^u} \equiv -\frac{\mathbf{tr}(\mathbf{J}\log\mathbf{J})}{H(X_\Omega, X'_\Omega)}, \qquad (34)$$

where $\mathbf{tr}(\mathbf{J}\log\mathbf{J})$ is the *trace* of the matrix $\mathbf{J}\log\mathbf{J}$, i.e.:

$$\mathbf{tr}(\mathbf{J}\log\mathbf{J}) = \sum_{i\leq n} J_{ii}\log J_{ii}. \qquad (35)$$

By dividing $\mathbf{tr}(\mathbf{J})$ by $H(X_\Omega, X'_\Omega)$ we capture the fraction of bits obtained from the observation of consistent pairs against all possible pairs $\langle m_i, m_j\rangle$[2].

The amount of *Consistent Information*, $\mathcal{I}(A^v \to A^u)$, is obtained by weighting the overall mutual information with the referential parameter:

$$\mathcal{I}(A^v \to A^u) = I(A^v \to A^u)\sigma_{A^v\to A^u}. \qquad (36)$$

The average of consistent information among two agents, $\mathcal{F}(A^v, A^u)$ will be, consistently:

$$\mathcal{F}(A^v, A^u) \equiv \frac{1}{2}\left(\mathcal{I}(A^v \to A^u) + \mathcal{I}(A^v \to A^u)\right). \qquad (37)$$

Since $\sigma_{A^v\to A^u} \in [0, 1]$, from the defintion of channel capacity and the symmetry properties of the mutual information, it is straightforward to show that:

$$\mathcal{F}(A^v, A^u) \leq \langle I(A^v, A^u)\rangle \leq \mathcal{C}(\Lambda).$$

Eqs. (34, 36) and (37) are the central equations of this paper. Let us focus on eq. (36). In this equation, we derive the average of consistent bits in a minimal system consisting of two agents (coder/decoder). Consistent information has been obtained by mathematically inserting the dual nature of the communicative sign - which forces the explicit presence of coder, channel and decoder modules- and subsequently selecting the subset of correlations by which the input symbol (the specific realization of X_Ω) is equal to the output symbol (i.e., the specific realization of X'_Ω). Eq. (37) accounts for the (possibly) symmetrical nature of the communicative exchange among agents: a priori, all agents can be both coder and decoder, and we have to evaluate and average the two possible configurations. The information-theoretic flavour of $\mathcal{F}$ enables us to study the conservation of referentiality from the well-grounded framework of Information Theory.

B. General Behavior of Consistent Information

So far we have been concerned with the derivation of the amount of information which is consistently decoded,

[2] We might notice that the amount of information carried by consistent pairs resembles the formal exposition of the Von Neumann entropy for quantum states, $S(\rho)$, which captures the degree of mixture of a given quantum state and its associated uncertainty in measuring (Von Neumann, 1936). In this way, we observe that S can be, roughly speaking, identified with an indicator of the consistency of the quantum state. However, it is worth noting that these measures are conceptually and formally different.

taking into account the dual nature of the communicative sign -equations (34), (36) and (37). Now we explore some of its properties, and we highlight the conceptual and quantitative differences between $\mathcal{I}$ and I.

To study the behavior of $\mathcal{I}$ and its relation to I, we will isolate the first three most salient features. Specifically, we shall concern ourselves with the following logical implications:

$$i)\ (\sigma_{A^v\to A^u}=1)\ \Rightarrow\ (I(A^v\to A^u)=H(X_\Omega)), \quad (38)$$
$$ii)\ (\sigma_{A^v\to A^u}=1)\ \not\Leftarrow\ (I(A^v\to A^u)=H(X_\Omega)). \quad (39)$$

The first $i)$ implication refers to the perfect conservation of referentiality, which, in turn, implies maximum mutual information. However, the inverse, $ii)$, is not generally true, since, as we shall see, there are many situations by which the mutual information is maximum although there is no conservation of referentiality. Furthermore, we consider a third case, the noisy channel (which implies that $H(X_\Omega|X'_\Omega)>0$). In this case:

$$iii)\ H(X_\Omega)>I(A^v\to A^u)>\mathcal{I}(A^v\to A^u). \quad (40)$$

We begin with the implications $i)$ and $ii)$. In both cases, the whole process is noiseless, since from eq. (27) $\max I(A^v\to A^u)=H(X_\Omega))$. To address the first logical implication, $i)$, we obtain the typology of configurations of $\mathbf{P}^v,\Lambda,\mathbf{Q}^v$ leading to $\sigma_{A^v\to A^u}=1$. We observe that the condition (39) is achieved if $\mathbb{P}(X'_\Omega|X_\Omega)=\mathbb{1}$, i.e., the identity matrix:

$$\mathbb{1}_{ij}=\begin{cases}1 \text{ iff } i=j\\ 0 \text{ otherwise.}\end{cases} \quad (41)$$

Such a condition only holds if

$$\mathbf{P}^v=(\Lambda\mathbf{Q}^u)^{-1}, \quad (42)$$

since given a square matrix $\mathbf{A}$, $\mathbf{A}\cdot\mathbf{A}^{-1}=\mathbb{1}$ -provided that $\mathbf{A}^{-1}$ exists. From the conditions imposed over the transition matrices provided in eqs. (6,12,17), the above relation is fullfilled if and only if all the matrices $\mathbf{P}^v,\Lambda,\mathbf{Q}^u$ are *permutation matrices*. Let us briefly revise this concept, which will be useful in the following lines. A *permutation matrix* is a square matrix which has exactly one entry equal to 1 in each row and each column and 0's elsewhere. For example, if $n=3$, we have 6 permutation matrices, namely:

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix},\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}, \quad (43)$$

$$\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix},\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}. \quad (44)$$

The set of $n\times n$ permutation matrices is idicated as $\Pi_{n\times n}$ and it can be shown that, if $\mathbf{A}\in\Pi_{n\times n}$, $\mathbf{A}^{-1}=\mathbf{A}^T\in\Pi_{n\times n}$ and, if $\mathbf{A},\mathbf{B}\in\Pi_{n\times n}$, the product $\mathbf{AB}\in\Pi_{n\times n}$. Furthermore, it is clear that $\mathbb{1}\in\Pi_{n\times n}$. If we translate the above facts of permutation matrices to our problem, we find that $\sigma_{A^v\to A^u}=1$ is achieved if:

$$(\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n}) \text{ and } \mathbf{P}^v=(\Lambda\mathbf{Q}^u)^T, \quad (45)$$

leading to the following chain of equalitites, which only holds in this special case:

$$\begin{aligned}\mathcal{I}(A^v\to A^u)&=I(A^v\to A^u)\\&=\max I(A^v\to A^u)\\&=H(X_\Omega).\end{aligned}$$

Case $ii)$ is easily demonstrated by observing that, if $\mathbb{P}(X_\Omega|X'_\Omega)\in\Pi_{n\times n}$, then $\mathbb{P}(X'_\Omega|X_\Omega)\in\Pi_{n\times n}$ and thus

$$H(X_\Omega|X'_\Omega)=0, \quad (46)$$

leading to:

$$I(A^v\to A^u)=\max I(A^v\to A^u)=H(X_\Omega), \quad (47)$$

which is achieved only imposing that

$$\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n}. \quad (48)$$

However, as we saw above, only a special configuration of permutation matrices leads to $\sigma_{A^v\to A^u}=1$. Thus, for the majority of cases where $I(A^v\to A^u)=\max I(A^v\to A^u)$, the conservation of the referentiality fails, leading to

$$I(A^v\to A^u)>\mathcal{I}(A^v\to A^u), \quad (49)$$

unless condition (45) is satisfied. Let us notice that there are limit cases where, although $I(A^v\to A^u)=\max I(A^v\to A^u)$, $\mathcal{I}(A^v\to A^u)=0$, since it is possible to find a configuration of $\mathbf{P}^v,\Lambda,\mathbf{Q}^u\in\Pi_{n\times n}$ such that $\mathbb{P}(X_\Omega|X'_\Omega)$ is a permutation matrix with all zeros in the main diagonal, leading to $\sigma_{A^v\to A^u}=0$.

Case $iii)$ is by far the most interesting, since natural systems are noisy, and the conclusion could invalidate some results concerning the information measures related to systems where referentiality is important. The first inequality trivially derives from equation (21), from which we conclude that $I(A^v\to A^u)<H(X_\Omega)$. The argument to demonstrate the second inequality lies on the following implication:

$$(H(X_\Omega|X'_\Omega)>0)\Rightarrow(\mathbb{P}_{A^v\to A^u}(X'_\Omega|X_\Omega)\notin\Pi_{n\times n}). \quad (50)$$

Indeed, let us proceed by contradiction: Let us suppose that $\mathbb{P}_{A^v\to A^u}(X'_\Omega|X_\Omega)\in\Pi_{n\times n}$. Then, as discussed above, $\mathbb{P}_{A^v\to A^u}(X_\Omega|X'_\Omega)\in\Pi_{n\times n}$. But this should imply that $H(X_\Omega|X'_\Omega)=0$, thus contradicting the premise that $H(X_\Omega|X'_\Omega)>0$.

This has a direct consequence. Since such conditional probabilities satisfy eq. (17), then, more than n matrix elements of $\mathbb{P}_{A^v\to A^u}(X_\Omega|X'_\Omega)$ must be different from zero. The same applies to the matrix of joint probabilities $\mathbf{J}$

and thus it also applies to $-\mathbf{J}\log\mathbf{J}$. Since the trace is a sum of n elements, it should be clear that, under noise:

$$H(X_\Omega, X'_\Omega) > -\mathbf{tr}(\mathbf{J}\log\mathbf{J}), \tag{51}$$

leading to:

$$\sigma_{A^v\to A^u} < 1, \tag{52}$$

thus recovering the chain of inequalities provided in eq. (40):

$$H(X_\Omega) > I(A^v \to A^u) > \mathcal{I}(A^v \to A^u). \tag{53}$$

If we expand the reasoning to the symmetrical consistent information $\mathcal{F}(A^v, A^u)$ defined in (37):

$$\mathcal{F}(A^v, A^u) < \langle I(A^v, A^u)\rangle. \tag{54}$$

We see that referentiality conservation introduces an extra source of dissipation of information. In those scenarios where referentiality conservation is an important advantage, the dissipation of information, $\mathcal{I}_D$, among two agents has two components:

$$\mathcal{I}_D = \overbrace{H(X_\Omega|X'_\Omega)}^{physical\ noise} + \overbrace{(1-\sigma)I(A^v \to A^u)}^{Referential\ noise}, \tag{55}$$

being the amount of useful information provided by consistent information, namely:

$$\mathcal{I}(A^v \to A^u) = H(X_\Omega) - \mathcal{I}_D. \tag{56}$$

IV. CASE STUDY: THE BINARY SYMMETRIC CHANNEL

As an illustration of our general formalism, let us consider the standard example of a binary symmetric channel where we have two agents, A^v, A^u, sharing a world with two events, namely $\Omega = \{m_1, m_2\}$ such that $\mu(m_1) = \mu(m_2) = 1/2$.

Case 1: Non-preservation of referentiality. We will consider a case where $I(A^v \to A^u) = \max I$ but $\sigma_{A^v\to A^u} = \mathcal{I}(A^v \to A^u) = 0$. The transition matrices of agents A^v and A^u are identical and defined as:

$$A^{v,u} = \left\{\mathbf{P}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{Q}^{v,u} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right\}. \tag{57}$$

The channel between such agents, Λ, is noiseless:

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{58}$$

We begin by identifying the different elements involved in the process. First, from eq. (14) we obtain:

$$\mu'(m_1) = \mu'(m_2) = \frac{1}{2}.$$

The matrix of joint probabilities, $\mathbf{J}$, is -see eq. (33):

$$\mathbf{J} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}. \tag{59}$$

Thus, rearranging terms, the mutual information from A^u to A^v -see (eq. 28)- will be:

$$I(A^v \to A^u) = \log 2 = 1\ \text{ bit}.$$

We observe that, for a communication system consisting of two possible signals,

$$\max I = \log 2 = 1 \text{ bit}. \tag{60}$$

Thus the mutual information is maximum. However, it is evident that such a system does not preserve referentiality, since, if $X_\Omega = m_1$, then $X'_\Omega = m_2$, and viceversa. Indeed, let us first obtain the matrix $-\mathbf{J}\log\mathbf{J}$, which will be:

$$-\mathbf{J}\log\mathbf{J} = \begin{pmatrix} 0 & -\frac{1}{2}\log\frac{1}{2} \\ -\frac{1}{2}\log\frac{1}{2} & 0 \end{pmatrix}. \tag{61}$$

And, thus, by its definition, the referential term will be (eq. 34):

$$\sigma_{A^v\to A^u} = -\frac{\mathbf{tr}(\mathbf{J}\log\mathbf{J})}{\log 2} = 0, \tag{62}$$

(notice that $\log 2 = 1$, although we keep the logarithm for the sake of clarity) being the amount of consistent information:

$$\mathcal{I}(A^v \to A^u) = 0 \text{ bits}. \tag{63}$$

This extreme case dramatically illustrates the non-trivial relation between $\mathcal{I}$ and I, proposing a situation where the communication system is completely useless, although the mutual information between the random variables depicting the input and the output is maximum.

Case 2: Preservation of the referentiality. In this configuration, the referentiality is conserved. Let us suppose a different configuration of the agents. Now the transition matrices of agents A^v and A^u are identical and defined as:

$$A^{v,u} = \left\{\mathbf{P}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{Q}^{v,u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\}. \tag{64}$$

The channel between such agents, Λ, is the two-dimensional noiseless channel defined in eq. (58). It is straightforward to check that the mutual information is maximal (= 1 bit), as above. The matrix $-\mathbf{J}\log\mathbf{J}$ will be, now,

$$-\mathbf{J}\log\mathbf{J} = \begin{pmatrix} -\frac{1}{2}\log\frac{1}{2} & 0 \\ 0 & -\frac{1}{2}\log\frac{1}{2} \end{pmatrix}. \tag{65}$$

This leads to $\sigma_{A^v\to A^u} = 1$, and, consequently:

$$I(A^v \to A^u) = \mathcal{I}(A^v \to A^u). \tag{66}$$

The above configuration is the only one which leads to $I = \mathcal{I}$. Furthermore -as shown in section III.b- it can only be achieved when I is maximum, i.e., in a noiseless scenario. In the last example we will deal with a noisy situation.

Case 3: Noisy channel. We finally explore the case where the matrix configuration of agents is the same as in the above example (eq. 57) but the channel is noisy, namely:

$$\Lambda = \left(\begin{array}{cc} 0.9 & 0.1 \\ 0.1 & 0.9 \end{array} \right). \tag{67}$$

We first derive the matrix of joint probabilities, $\mathbf{J}$,which takes the following form:

$$\mathbf{J} = \left(\begin{array}{cc} 0.45 & 0.05 \\ 0.05 & 0.45 \end{array} \right). \tag{68}$$

We now proceed by observing that $\mu'(m_1) = \mu'(m_2) = 1/2$. Thus, the mutual information will be:

$$\begin{aligned} I(A^v \to A^u) &= 0.9 \log \frac{0.45}{0.5 \cdot 0.5} + 0.10 \log \frac{0.05}{0.5 \cdot 0.5} \\ &= 0.531... \text{ bits.} \end{aligned} \tag{69}$$

To evaluate the degree of consistency of the communicative system, we firstly compute the matrix $-\mathbf{J}\log\mathbf{J}$:

$$\left(\begin{array}{cc} -0.45\log 0.45 & -0.05\log 0.05 \\ -0.05\log 0.05 & -0.45\log 0.45 \end{array} \right) = \left(\begin{array}{cc} 0.518 & 0.216 \\ 0.216 & 0.518 \end{array} \right). \tag{70}$$

Since $H(X_\Omega, X'_\Omega) = 1.468$ bits, the referential parameter is:

$$\begin{aligned} \sigma_{A^v \to A^u} &= -\frac{\mathbf{tr}(\mathbf{J}\log\mathbf{J})}{H(X_\Omega, X'_\Omega)} \\ &= \frac{0.518 + 0.518}{1.468} \\ &= 0.706... \text{ consistent bits/bit.} \end{aligned} \tag{71}$$

(where the last "bit" refers to "bit obtained from the observation of input-output pairs"). The consistent information is, thus:

$$\begin{aligned} \mathcal{I}(A^v \to A^u) &= I(A^v \to A^u)\sigma_{A^v \to A^u} \\ &= 0.531 \times 0.706 \\ &= 0.375 \text{ bits.} \end{aligned} \tag{72}$$

Due to the symmetry of the problem, the average among the two agents is:

$$\mathcal{F}(A^v, A^u) = 0.375 \text{ bits.} \tag{73}$$

The amount of dissipated information is, thus:

$$\mathcal{I}_D = \overbrace{0.469}^{physical\ noise} + \overbrace{0.156}^{Referential\ noise} \text{ bits.} \tag{74}$$

We want to stress the following point: The matrix configuration is consistent with the framework proposed in *case 2*, where the amount of consistent information is maximum, but now the channel is noisy. The noisy channel has a double effect: first, it destroys information in the standard sense, since the noise parameter $H(X_\Omega|X'_\Omega) > 0$, but it also has an impact on the consistency of the process, introducing an amount of referential *noise* due to the lack of consistency derived from it. Thus, as derived in section III.b, eq. (40), in the presence of noise, we have shown that the inequalities

$$H(X_\Omega) > I(A^v \to A^u) > \mathcal{I}(A^v \to A^u) \tag{75}$$

hold, being, in our special case:

$$1 > 0.531 > 0.375. \tag{76}$$

V. DISCUSSION

The accurate definition of the amount of information carried by consistent input/output pairs is an important component of information transfer in biological or artificial communicating systems. In this paper we explore the central role of information exchanges in selective scenarios, highlighting the importance of the referential value of the communicative sign.

The conceptual novelty surrounding the paper can be easily understood from the role we attribute to *noise*. Physical information considers a source of $H(X)$ bits and a *dissipation* of $H(X|Y)$ bits due to, for example, thermal fluctuations. We add another source of information dissipation: the non-consistency of the pair signal/referent, putting aside the degree of correlation among random variables (see eq. 55). Indeed, in many physical processes no referentiality is at work, perhaps because, it is not relevant to wonder about the consistency of the communicative process. Moreover, if the whole system is *designed*, consistency problems are apriori ruled out, unless the engineer wants to explicitly introduce disturbances in the system. What makes biology different, however, is that biological systems are not designed but instead, are the outcomes of an evolutionary process where the nature of the response to a given stimulus is important, which makes the problem of consistency relevant for evolutionary scenarios. This problem needs an explicit formulation, being what we called *consistent information* the theoretical object that links raw information and function, or environmental response.

Are information processing mechanisms of living systems optimal regarding referentiality conservation? As we discussed above, it seems reasonable to assume that the conservation of referentiality must be at the core of any communicative system with some selective advantage. The general problem to find the optimal code, however, resembles the problem of finding the channel capacity, for which is well known that no general procedure exists (Cover and Thomas, 1991). Thus, how autonomous systems deal with such a huge mathematical

problem? One may consider the possibility that the coevolution of the abstract coding and decoding entities; this would avoid the system to face a great amount of configurations per generation, thereby being all options highly limited at each generation where selection is at work.

We finally emphasize that the unavoidable dissipation of mutual information points to a reinterpretation of information-transfer phenomena in biological or self-organized systems, due to the important consequences that can be derived from it. Further work should explore the relevance of this limitation on more realistic scenarios, together with other implications that can be derived by placing equation (36) at the center of information transfer in biology.

Acknowledgments

We thank the members of the Complex Systems Lab for useful discussions. This work has been supported by a Juan de la Cierva grant from te Ministerio de Ciencia y Tecnología (JF), the James S. McDonnell Foundation (BCM) and by Santa Fe Institute (RS).

References

Ash, R. B., 1990, *Information Theory* (New York. Dover).

Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory* (John Wiley and Sons. New York).

Haken, H., 1978, *Synergetics: An Introduction. Nonequilibrium Phase Transitions and Self- Organization in Physics, Chemistry and Biology (Springer Series in Synergetics)* (Springer).

Hopfield, J., 1994, Journal of Theoretical Biology **171**(1), 53,

Hurford, J., 1989, Lingua **77**(2), 187,

Kolmogorov, A., 1965, Problems Inform. Transmission **1**, 1.

Komarova, N. L., and P. Niyogi, 2004, Art. Int. **154**(1-2), 1,

Ming, L., and P. Vitányi, 1997, *An introduction to Kolmogorov complexity and its applications* (Springer, New York [u.a.]),

Niyogi, P., 2006, *The Computational Nature of Language Learning and Evolution* (MIT Press. Cambridge, Mass.),

Nolfi, S., and M. Mirolli, 2010, *Evolution of Communication and Language in Embodied Agents* (Berlin. Springer Verlag),

Nowak, M. A., 2000, Philosophical Transactions: Biological Sciences **355**(1403), 1615,

Nowak, M. A., and D. Krakauer, 1999, Proc. Nat. Acad. Sci. USA **96**(14), 8028,

Plotkin, J. B., and M. A. Nowak, 2000, Journal of Theoretical Biology **205**(1), 147,

Saussure, F., 1916, *Cours de Linguistique Générale* (Bibliothèque scientifique Payot: Paris),

Shannon, C. E., 1948, Bell System Technical Journal **27**, 379.

Steels, L., 2001, IEEE Intelligent Systems **16**, 16, ISSN 1541-1672.

Steels, L., and J.-C. Baillie, 2003, Robotics and Autonomous Systems **43**(2-3), 163.

Von Neumann, J., 1936, *The Mathematical Foundations of Quantum Mechanics* (Princeton University Press. Princeton, N. J.).

Yockey, H. P., 1992, *Information Thery and Molecular Biology* (Cambridge University Press, Cambridge (UK)).